VQEG Meeting Minutes

**Mountain View, California, USA**

**November 12 to 16, 2018**

**Remote participation link:** [**https://meet.google.com/hip-yutw-bjr**](https://meet.google.com/hip-yutw-bjr)

Day 1 – Monday, Nov 12

Group Reports

**AVHD-AS** mostly working on model validation with ITU-T SG12, Q14, project AVHD-AS/P.NATS Phase II. Model submission is Dec. 14, 2018 due to delays. Models will be validated on short videos (6-8 s) and long videos, depending on model type.

**PsyPhyQA** identified source videos for experiment, conducted cross lab comparisons, and are ready to run an experiment. PsyPhyQA welcomes more organization participation. Would like to have 1 or 2 more labs to validate this experiment.

**eLetter** call for papers on PsyPhyQA went out recently. Goal is a draft eLetter by January.

**SAM** has bi-weekly teleconferences. Results will be presented this week. Have proposal for data format from subjective experiments. Have a few solutions for problems identified at the last VQEG meeting.

**CGI** Started at the last meeting with an expression of interest in computer generated imagery, gaming content quality. Seeking interested parties.

**NORM** Discussions on what a no reference metric should be like: what is the scope, what values should it predict, what problems should it detect.

**JEG-Hybrid** open collaboration for robust hybrid perceptual model. Have robust database with simulated packet loss, etc. Looking for systematic weaknesses of objective metrics.

**QACoViA**  Continuation of the QART project, with an expanded and shifted scope.

**IMG** quality assessment for immersive media, 360 video, etc. Monthly audio calls, seeking specific use cases for collaborative work. Lab equipment identified. Working on test plan on Friday. Working collaboratively with ITU-T SG12, Q13, on draft Recs. related to immersive media quality assessment.

**HFVE** Established at last meeting. Focus on IEEE P333.1 item, which has established some standards already.

**Joint Qualinet-VQEG Team on Immersive Media**, goal to inform each of these groups of the other’s activities. Definition of QoE from QUALINET has in fact been standardized in ITU-T P.10

**Tools and Subjective Labs Setup** Have website collection of tools. Migrated to a new platform. If have tools for the community, reach out to this group.

**VQEG Administration and Web Suppor**t Reflectors continue to be a problem. No solution currently works for everyone. SAM has moved to Google Groups.

**MPEG** Working on point-cloud compression and neural network compression. Call for proposals on video compression beyond HEVC, SDR / HDR / 360 degree video.

**ITU-T SG12** Mentioned above (AVHD-AS/P.NATS Phase II) Also working on speech quality measurements, telepresence, P.QUITS initial buffering / rebuffering before people quit; how to perform subjective tests to detect the likelihood of people quitting, VR and 360 degree video (breakout today, and to be discussed later this week), , .

**Alliance for Open Media** AV1 has been  (late June / July); work on next version progressing; intense work on productizing, Microsoft indicates Windows / Chrome support, hardware being developed

**ICDM** Chair will come Wednesday to give presentation; working on next version of display standard, want to update chapter 4, visual inspection. ICDM is reaching out to VQEG for help with this update.

Presentations

Werner Robitza, “New Tools and Subjective Labs Website”

* New website based on GitHub to list available software, tools, and database.
* <https://vqeg.github.io/software-tools/>
* Look at the site, let Glenn and Werner know about new tools
* There is interest in merging the Qualinet database (http://dbq.multimediatech.cz/) into this website or a similar setup

**Action:** the group chairs will propose a collaborative VQEG / Qualinet effort to merge these two websites. A liaison will be sent from VQEG to Qualinet about it. Werner and Lukas will make a first draft

Computer Graphics (CGI)

Steven Schmidt, “Ongoing standardization activities of gaming quality of experience”

* ITU-T Rec. G.1032
* ITU-T Rec. P.809, includes subjective test methods (interactive vs passive)
* We are always searching for collaborations
* [www.qu.tu-berlin.de](http://www.qu.tu-berlin.de) for more information

Saeed Shafiee Sabet, “Gaming Video Datasets”

* Gaming Video dataset (KU-DT/TUB), KUGVD (KU), and CGVDS (DT/TUB-Simula)
* Dataset distributed soon
* Investigated usage rights from game owners: this use and redistribution is allowed

Saman Zadtootaghaj, “Performance evaluation of existing quality models and ITU standards on video gaming quality estimation”

* Cloud gaming: diverse in video complexity
* Existing metrics rarely trained no this content

No Reference (NR) Metrics  (NORM)

Mikołaj Leszczuk, “Quality Indicator - Lip Sync”

* Summary of lip sync influence, taken from various reports
* Metric to detect lip sync problems, through comparison of audio and video

Yilin Wang, “Quality analysis for user generated contents”

* Many videos uploaded to YouTube are not pristine, so full reference metrics do not work
* Described several NR metrics from published papers; references given
* Idea: report scores for each artifact + overall quality assessment

Nabajeet Barman, “ML based no reference gaming video quality estimation”

* Passive gaming video streaming
* Three models: NR-GVSQI, NR-GVSQE, NR-GVQM
* Datasets GamingVideoSet (available open source) and KUGVD
* Three different approaches to machine learning for NR metric design

Margaret Pinson, “What should an NR metric scatter plot look like?”

* Mylene Farias (Unb): Scores of different FR and NR metrics vary (in a good scenario, are scaled) for different datasets with different spatial and temporal resolutions.  In one previous work by our group, we tested the performance of metrics to these variations and adapted the videos (automatically) to obtain the best performance. Our goal in this case was actually to reduce both these dimensions to reduce computational complexity, but the same idea can be used in the design of NR metrics, since the same parameters are also available. One interesting point from this work is that the effect of spatial and temporal resolutions depends on the type of distortion, which means that a classification of the type of distortion is certainly useful.

(For details, see:  <https://www.researchgate.net/publication/327828616_A_framework_for_computationally_efficient_video_quality_assessment>)

* Lukas Krasula (Arm): The change in performance of particular sub-metrics could potentially be resolved by their fusion using a locally adaptive system, as proposed by Barri et al. (<https://ieeexplore.ieee.org/document/6786302>).
* Lukas Krasula (Arm): The task of evaluating quality of images/videos that are presented to observers only once is very challenging and will most probably result in high variance of observer scores. It is therefore dangerous to take MOS without considering their respective standard deviations. I would recommend the use of the performance evaluation methodology presented in here: <https://ieeexplore.ieee.org/document/7498936>. It not only considers standard deviations of the subjective scores but also enables meaningful combination of data coming from multiple databases without introducing noise caused by different quality scales among the experiments conducted when creating them. I am willing to provide any necessary assistance when processing the data using this method.

Discussion on No Reference Metrics

Where do we want to go? What do we want to accomplish?

Facebook; way of measuring user created content; willing to help with content generation

Limited to human consumption of video (computer use is separate project)

Professional & consumer generated content both of interest

Standardization only useful across industry (do not need internally); measurement standard

Sky UK: professional content, confidence score for the measurement (e.g., 4 mos +/- 1)

AT&T/DirecTV: professional content, contribution quality, distribution quality,

We can’t make it better but can make it much worse;

If incoming quality is below a threshold, then reject (if have standard scale, company can set a threshold)

AGH: What is quality? We don’t have a strict definition. This would help us limit scope (this part does not interest me vs this part does)

Map back to VMAF? Or include satisfaction, aesthetics, etc.?

Need a definition or example of the best possible quality.

QoE is slightly different from video quality. Enjoy watching pets even if poor quality. QoE includes audio quality, and other factors. If restrict to just video, still not one size fits all (e.g., live video vs gaming).

White paper / open metric vs standard?

* NR is too open / vague right now for an ITU Recommendation. ITU-T SG12 is discussing how to treat machine learning models.
* Some skepticism expressed on standardizing a list of numbers (machine learning)
* Qualinet white paper on QoE, see ITU-T Rec P.10. However, also some abuse exists.
* Separate standard from development work

Bar is low. Typical consumer equates resolution to quality.

To have impact, we must target a business problem, business decision

Artistic quality probably must be left out.

Standardize our viewing conditions. Device has impact, both in development & validation. Make these high level, practical decisions.

Original goal: toolsets to enable industry decisions. Consumers have difficulty understanding MOS from metrics (PSNR, VMAF). Need application specific tools that engineers understand and can use, including “**what is a meaningful difference**” for their specific application. Engineers don’t understand the numbers from our metrics.

JND is a solution to above, mentioned at a prior meeting.

**Decided:**

* Will start monthly calls.
* Goal of NORM:  Define a No-Reference metric that can accurately measure the visual quality of the decoded video “signal.”  Has at least a Mean Opinion Score (MOS) as a metric.  Has a specified just-noticeable-difference (JND) and confidence interval.  Has guidance on how to interpret the results.

**Open Actions:**

* Discuss MOS and JND on first call.
* Do we combine User Generated and Camera Capture?

**Professional Content:** Sky UK, Spirent, Facebook, AT&T/DirecTV, YouTube, Kingston Univ.

**User Generated Content:** YouTube, Facebook, Spirent, ByteDance, NTIA/ITS, Kingston Univ.

**Camera Capture:** NTIA/ITS (This may be merged with user generated content)

Day 2 – Tuesday, Nov 13

SAM (Statistical Analysis Methods)

Intro (Lucjan). SAM is working on:

* Subject model (tbd on Thursday)
* Other stuff: statistical analysis in general, format to report subjective tests, tc.

#104, Kjell Brunnström, “Statistical quality of experience analysis”

* Link to original article: <https://www.spiedigitallibrary.org/journals/journal-of-electronic-imaging/volume-27/issue-05/053013/Statistical-quality-of-experience-analysis--on-planning-the-sample/10.1117/1.JEI.27.5.053013.full>
* Within-subject design (typical scenario in subjective tests is that each user scores several PVSs): dependent T-test for paired samples.
* Estimation of sample size for statistical power (e.g. 0.8), MOS difference (0.5, 1.0), standard deviation (0.8), number of comparisons.



* Prepared input to ITU SG12: amend P.1401 with a text on multiple comparison, harmonization of texts in BT.500, P.910, P.913.
	+ No pre-planned number of comparisons: 27 in controlled env, 37 in public env.
	+ Pre-planned number on comparisons (<100): 18 and 25 subjects respectively.
* Comments about the proposal:
	+ Patrick: “magic” numbers in standards may be misleading; it’s better to propose tools
	+ Ioannis: other statistical analysis tools (e.g. removing user bias) may reduce the standard deviation and thus the number of subjects needed
	+ Further discussion on the topic (Margaret, Lucjan, Ioannis, Patrick).

#103, Lucjan Janoswki, [“JSON for Subjective Experiments”](https://docs.google.com/presentation/d/1QwbjwIv0mgbqor5EtARce1OvK0I_2UdOkS3dEVrGr1E/edit#slide=id.g47d2227892_3_49)

* Based on *sureal* package
* Per PVS, per subject, object-oriented
* Discussion on naming, etc. (several attendees)
* suJSON: a JSON for subjective tests’ data
* **Proposals:**
	+ **Add a field for pre- and post-experiment Q&A**
	+ **Add viewing conditions**
	+ **Prepare format for pair comparison (e.g. PVS\_id list at scores)**
	+ **Do not duplicate information (remove PVSs scored by subject in subject data, and subjects scoring PVSs in PVS data)**

IMG (Immersive Media Group)

#109, Francesca de Simone, “Measuring User Quality of Experience in Social VR Systems”

* User QoE in sVR has multiple dimensions (e.g. social presence, AV quality)
* Factors impacting QoE:
	+ Avatar appearance
	+ Avatar behavioural realism
* Experiment: view a movie trailer in pairs
	+ Compare: f2f, Facebook, TNO
	+ Measure:
		- Subjective: Quality of Interaction, Social Connectedness, Presence/Immersion
		- Objective: user behaviour (user talk to / look at each other)

#110, Ashutosh Singla, “Comparisons of Different Subjective Test Methods for HEVC Encoded Omnidirectional Videos”

* Measure user behaviour, video quality, simulator sequence
* Compare ACR, m-ACR, DSIS
* Results
	+ 4k vs 6k is observed, 6k vs 8k does not show any difference
	+ DSIS shows (slightly) better discriminative power than ACR
	+ Simulator sickness: increases along time (and with test duration), decreases with breaks, DSIS shows less sickness
* Discussion:
	+ Discussion on resolution of the display vs resolution of the video
* “TUIL 360º Video Dataset”
	+ 3x 8K 30-second high-quality video
* Preview of HVEI’19 presentation

#112, Pablo Perez, “Using 360 VR Video to Improve the Learning Experience in Veterinary Medicine University Degree”

* Preview of HVEI’19 presentation

QACoViA

#104, Mikołaj Leszczuk, “Evaluation of Multimedia Content Summarization Algorithms”

#105, Mikołaj Leszczuk, “Objective Video Quality Assessment Method for Recognition Tasks”

* Existing QoE objective metrics focus on “entertainment media”. They cannot predict the accuracy of recognition tasks with enough reliability.
* New project. Target: Develop a new objective video quality assessment method for recognition tasks.
* Evaluate against computer vision success instead of MOSs
* Comment (Patrick): What it is needed in QoCoVIA is establishing the statistical framework on how to test the quality estimator against the groundtruth. We also need a framework to “stress” the algorithm (which kinds of distortions / noise / … are applicable to CV use case).
* Spirent (Jeremy) is also interested in the topic. Today most data is transported uncompressed because of the fear that there might be issues if it is compressed. IEEE P2020 might be a group to collaborate with.
* Comment (Ioannis): is it possible to assign a score to a single signal (which will individually success or fail)?
* Mylene: should we enter into detecting images that can fool the algorithm? Patrick: we should not.

Immersive Media Group

#115, Kjell Brunnström, “Quality of Experience of Visual-Haptic interaction in a Virtual Reality simulator”

* Control of a truck crane with VR
* Test joystick delay in VR simulator
	+ 0 - 800 ms
* 31 test subjects (22m/9f), 2 didn’t finish the test due to sickness
* Results
	+ 0-400 ms no effect. 800 ms has significant effect in task completion (also subjectively).
* Next steps:
	+ Evaluate with experts (crane operators)
	+ Test with a remote controlled miniature crane, and also with the actual crane.

#129, Neil Birkbeck, Roberto Azevedo, “Understanding distortions in Youtube omnidirectional videos”

* Artifact ranking:
	+ Blur, blockiness, seams, h264 qp
	+ Focus: 1st JND and immersion breaking point
* Testing procedure:
	+ Find JND (user add degradation until notice difference with original)
	+ User add degradation until “immersion” is broken.
	+ No time limit per user
* Results
	+ Artifacts tend to be more annoying (break immersion) in stereo than in mono
	+ Existing metrics do not quantify localized artifacts (e.g. seams)
	+ Blurring seems to be less perceptible than blocking

JEG Hybrid

#108, Enrico Masala, “Objective Measures on the ITS4S Database”

* Objectively-annotated Large Scale Database
* Subjective annotation of ITS4S database
	+ ITS4S already contains subjectively annotated data
	+ Comparison with objective metrics (mostly VMAF)
* Data available for download

Florence Agboma, “Characterising Objective Metrics Using Large-Scale Database” (discussion session)

* Scope: Compare open source with commercial tools (objective metrics)
* Look for places where metrics agree and put the ground truth there (and do the subjective tests there).
* Understand weaknesses → better interpret the results

SAM / IMG

Lucjan Janowski, “Repetition button”

* Button where users report that they were unable to rate the sequence (they were distracted)
* Two options: repeat button vs skip (discard that subject’s score for one PVS)
* SAMVIQ allows re-play of videos
* Medical doctors allowed to replay because they are used to repeatedly viewing sequences
* Comparison tests with small differences are improved by repeat option
* Label? Maybe “I missed the sequence”
* Should note who used the button, for which sequences, who used the button often, …
* Self reporting & discard (skip) will have less impact on ACR task
* Could repeat at a random time later in the session (i.e., discard this score, come back to later)
* VQEG should consider using experts in subjective tests who can more consistently and accurately score video than consumers.

Day 3 – Wednesday, Nov 14

New Group Proposal

Proposal for new VQEG group: 5GKPI presented by Pablo    Perez

* Relationship between 5G key performance indicators and video QoE
* Proposal for starting group, in the future the group will propose test plans

Discussion:

* Is the group limited to 5G or any new network technologies?
	+ it is a good suggestion
* What part of 5G we are talking about? Is it the first step to answer this question by finding where are the problems?
	+ We start from entertainment video and self-driving cars.
* SG12 already decided to work on 5G. The work should be synchronized.
* Who is interested?
	+ NTIA/ITS,
	+ University of West Scotland,
	+ University of Madrid,
	+ Intel,
	+ Facebook,
	+ Opticom,
	+ Spirent,
	+ University of Nantes,
	+ Vittorio,
	+ Kingston University,
	+ RISE (Acreo)
	+ Nokia
* SG12 discussions, exchange of ideas would be helpful

IMG

Pablo provided summary of what was done.

Presentation: Narciso García    “Video Multimethod Assessment Fusion (VMAF) on 360VR”

*Discussion and Questions:*

- What is the input format for the VMAF

   - it is the rectangular form

- what was the subject behavior

  - they mostly watched equator

- It is surprising that VMAF works so well, for pools some distortions, not visible for VMAF should distort.

- Maybe the detail of the algorithm is limiting the pools influence

- only the same device?

  - yes

- how voting was made?

  - within an application by pressing a button on a headset

- Maybe you bias the chosen video by getting VMAF results before the experiment. There was a long discussion if the experiment design influenced the results. There are different ideas who it influenced the results.

- Did you use a different metric? If PSNR works as well then we know that it is not proved that VMAF is working so well.

  - It is planned but not yet done

- The results show that influence on quality almost does not depend on the content, which is surprising.

ITU-T Q13 Reporting

Kazuhisa Yamagishi

* What is the difference between G.QoE-VR  and P.360-VR
	+ G.QoE-VR describes only factors influencing and P.360-VR describes the test methodology.

Showing video impairments which should be detected by NR metrics produced by Sky

By Florence Agboma

IMG

Francesca De Simone “Evaluation of immersion and presence in VR”

*Discussion and Questions:*

* Francesca proposed to start collaborative work on the measuring and defining presence.
* How can we measure the time duration presence
	+ No ready solution
* Is there any other people interested:
	+ There are some people interested.
* What is the main reason to understand presence, will you prepare sequence which fit more to specific presence level?
	+ This is part of QoE, that is why we have to measure

Pablo Pérez, Jesús Gutiérrez “Methodologies for subjective quality evaluation of short and long 360-degree videos”

    *Discussion and Question:*

* Do you know if 20 is not just exploring the surrounding but not "seeing" the details?
	+ we are still exploring, so any literature can be helpful.
* We do not show a long sequence because then the content become important.
	+ It is true that time duration and interest in the sequence start to be important.
* People attention after 12 minutes drops significantly
	+ It is true, it is one of the problems related to this methodology.
* What are we asking about? Do we have MOS?
	+ no MOS in most of the cases, we ask more detail questions.
* Can we divide long sequence too short and model the head movements as the separate influencing factor?
	+ yes, but it will not cover all the use-cases.
* Reaction time will have a strong influence on continues voting.
	+ yes, it is a problem
* Do you assume that we are using TCP type service? ITU did the evaluation with one minute long sequences.
	+ This is a separate problem - quality fluctuations.
* Long sequences voted also divided into shorter sequences. It helped with the better understanding of the user behavior.
	+ it is a good idea.

Jesús Gutiérrez “Monitoring user behavior in subjective tests with 360-degree video”

*Discussion and Question:*

* For any experiment head movement, data should be gathered.
* We can use the evaluated metrics to predict which content should be used for subjective experiments.

ICDM Presentation

Joe Miseli, “ICDM standard on visual assessment, an often hidden gem”

* Version 2 coming, [www.icdm-sid.org](http://www.icdm-sid.org)
* ICDM is making HDR test videos and images that will be make these freely available
* Opportunity to contribute subjective methods to assess display performance
* 6 month window

PsyPhyQA

Psycho-physiology quality assessment, Neem, Sebastian, and Ulrich; Maria very active

* Summary of main  objectives

#128, Naeem Ramzan, Stamos Katsigiannis, “A QoE and Simulator Sickness Evaluation of a Smart-Exercise-Bike Virtual Reality System via User Feedback and Physiological Signals”

* Assess exercise game’s VR display, when varying resolution and frame rate
* Visual quality and simulator sickness

PsyPhyQA test plan is available at <https://docs.google.com/document/d/1o52KrvHhUCdrs4lGjWKtDjkPWrcaOof0kOLLXdXfR7E/edit>

* Have subjective plan ready; seeking other labs interesting in running this test
* UWS, Fraunhofer, Kingston University, and maybe Univ of Nantes

The next **eLetter** will be on PsyPhyQA topic. Announcement sent to reflector.

QoMEX 2018 had a session on PsyPhyQA.

Book to be published soon: **AI for Emerging Verticals**, subtitle: Robotics, human computing interaction, sensing and networking

Future Meetings

* Will seek a host in China and/or co-loc SG12 interim meeting
* 1 year out, possibly Sky UK, UPM, UWS, Kingston University
* Amazon would like to host, maybe fall 2019, Seattle USA

Presentations

#117, Yasuko Sugito, “8K subjective evaluation experiments”

* Describes challenges from 8K resolution monitors

#128, Pablo Perez, “Key Performance Indicators of Video and 5G networks”

#136, Mylene Farias, “Analyzing the influence of cross-modal degradations on the perceived audio-visual quality”

Day 4 – Thursday, Nov 15

SAM

134, Lucjan Janowski, “Notation proposal”

    *Discussion and Question:*

* It is suggested to use sigma i, sigma j,... instead of upsilon, phi, …
There is strong rejection considering this variable in the future could not be standard deviation
* Suggestion to reconsider i, j, k as definitions and use other letters
Off line meeting will be held to decide

101, Lucjan Janowski, “Maximum Entropy Solution”

    *Discussion and Question:*

* Clarification is asked about the choice of maximum entropy

119, Krzysztof Rusek, presented by Lucjan Janowski , “Towards Bayesian Subject Model”

113, Jing Li, “Hybrid-MST: A Hybrid Active Sampling Strategy for Pairwise Comparison”

    *Discussion and Question:*

* It is commented that parallel execution could reduce complexity

111, Zhi Li, Analysis Tools in the VMAF Open-source Package

* VMAF includes tools to train and understand model performance, particularly better understanding of machine learning models trained to predict quality
* New techblog, VMAF: the journey continues: <https://medium.com/netflix-techblog/vmaf-the-journey-continues-44b51ee9ed12>
* Packet Video Workshop 2019, June in MA, USA

*Discussion and Question:*

* Additional features for analysis are proposed which should already be in VMAF but not exposed yet.
* It is commented that the bootstrapped results do not mean around the result of the full model. More research is needed to explain this behaviour.
* It is suggested to not restrict bootstrapping to video or subject, but to sampling any user vote of a PVS. This strategy was tested, but feedback from statistical experts indicated that the variation would then be underestimated.

Presentations

139, Ioannis Katsavounidis, “The many faces of Video at Facebook”

*Discussion and Question:*

* Are there HDR plans?
Yes, as soon as the content is delivered to Facebook

138, Derek Pang, “Video quality considerations for Android”

*Discussion and Question:*

* It is commented that standards like P.1203 already take into account the different viewing modes mobile and TV. Also with respect to test data, there is already a lot of different types of content used these days. Content type contributions are encouraged from anyone having specific content requirements.
* Did you ever do test equipment using a high end camera as reference?
Controlled labs are used to perform tests.
For control, full-reference metrics are indeed already used.
* Will you be sharing any of your no reference metrics?
At the moment no end-to-end reference metric and no plan to generate one.
* Can subjective test result be made available?
Will be considered.
* Analyzing the users cannot be done because of privacy, could you setup such a system using volunteers at massive scale?
This has been actively looked at.
* Have you seen evidence that localized artefacts influence opinion scores?
Depending on the situation, this indeed can be bad.
* Couldn’t you analyze the camera performance on a camera basis rather than a photo basis?
The video quality rather than the camera quality is investigated.
Although sensor noise and capture noise is the same, the processing happening inside can be very different.
* Aren’t these operations under control of Android?
Indeed, but the API developer can do a lot of tuning.

#142, Ioannis Katsavounidis, “Video codec comparison using the dynamic optimizer framework”

-     Opportunities using FR metrics to choose best resolution, QP, and codec (e.g., H.264 vs H.265 vs VP9)

#140, Narciso García, “Considerations on FTV quality assessment”

* FTV achieved in real time

#130, Ashutosh Singla, “Approaches for assessing the simulator sickness”

*Discussion and Question:*

* Difference from simulator and cyber sickness. Relation with immersive media?

3D depth needs to be calibrated

* Questionnaire not designed for this use case. Users might be biased. Probably simulator sickness questionnaire should not to be included in recommendations.

Physiological information should be included (information is noisy)

HFVE (Human Factors for Visual Experiences) Session

#137, Sanghoon Lee, “IEEE 3D human Factor WG: current status of P3333.1” (Presented by Maria Martini)

* Related to immersive content (not only 3D)

#135, Maria Martini, “Quality of Experience for Light Field imaging”

* Report on three studies: Impact on angular resolution, image resolution, and joint impact

Contribution by IMG?

PsyPhyQA and SAM could also be involved

**Decision**: Collect proposals and contributions by VQEG and open a line of communication with the IEEE standardization group (e.g. by somebody reporting in that group)

Day 5 – Friday, Nov 16

IMG

Work on test-plan, Pablo Perez

* Use cases: discussion, if entertainment video is going to be the mass application of 360 video?
* Watching 360 video sequences:
	+ Some shooting conditions are not well-suited for 360 video.
	+ Still difficult (unsuccessful) shooting footages should be kept in a dataset.
* More information in the following document: <https://docs.google.com/document/d/18L_Q7vyzEssoAzbRsTKBTtMrwfjISMf0HUGofPqOOJI>

#110, Ashutosh Singla, “Comparisons of Different Subjective Test Methods for HEVC Encoded Omnidirectional Videos” (fragments repeated, discussion on stitching artifacts)

Jesús    Gutiérrez, 360 content datasets

* More information in the following document: https://docs.google.com/document/d/1Q23TI2ARHWjP\_13Ngya9pp672CCo1WaK1wHI4eCCRc8/edit
* Discussion on subjective test methodology

Agreements:

* Test-plan for subjective test methodology to be developed
* Cross-lab common set to be created
* Test-plan by the next meeting
* Open questionnaire at the end of the test to capture open thoughts of subjects
* **Decision**: We will be testing:
	+ Length of sequences
	+ ACR vs DCR
	+ Influence of HMD
	+ Simulator sickness questionnaire or other relevant questionnaires
	+ Random viewports (To be decided)
* Collecting partners interested in conducting experiments:
	+ YouTube
	+ Gent
	+ RISE
	+ Evaltech
	+ UPM
	+ Ilmenau
	+ Nantes
	+ Nokia Bell Labs
	+ Kingston
* Interested partners from IMG meetings (not present)
	+ CWI
	+ Roma 3

VQEG Meeting Minutes

**Mountain View, California, USA**

**November 12 to 16, 2018**

**Remote participation link:** [**https://meet.google.com/hip-yutw-bjr**](https://meet.google.com/hip-yutw-bjr)

Day 1 – Monday, Nov 12

Group Reports

**AVHD-AS** mostly working on model validation with ITU-T SG12, Q14, project AVHD-AS/P.NATS Phase II. Model submission is Dec. 14, 2018 due to delays. Models will be validated on short videos (6-8 s) and long videos, depending on model type.

**PsyPhyQA** identified source videos for experiment, conducted cross lab comparisons, and are ready to run an experiment. PsyPhyQA welcomes more organization participation. Would like to have 1 or 2 more labs to validate this experiment.

**eLetter** call for papers on PsyPhyQA went out recently. Goal is a draft eLetter by January.

**SAM** has bi-weekly teleconferences. Results will be presented this week. Have proposal for data format from subjective experiments. Have a few solutions for problems identified at the last VQEG meeting.

**CGI** Started at the last meeting with an expression of interest in computer generated imagery, gaming content quality. Seeking interested parties.

**NORM** Discussions on what a no reference metric should be like: what is the scope, what values should it predict, what problems should it detect.

**JEG-Hybrid** open collaboration for robust hybrid perceptual model. Have robust database with simulated packet loss, etc. Looking for systematic weaknesses of objective metrics.

**QACoViA**  Continuation of the QART project, with an expanded and shifted scope.

**IMG** quality assessment for immersive media, 360 video, etc. Monthly audio calls, seeking specific use cases for collaborative work. Lab equipment identified. Working on test plan on Friday. Working collaboratively with ITU-T SG12, Q13, on draft Recs. related to immersive media quality assessment.

**HFVE** Established at last meeting. Focus on IEEE P333.1 item, which has established some standards already.

**Joint Qualinet-VQEG Team on Immersive Media**, goal to inform each of these groups of the other’s activities. Definition of QoE from QUALINET has in fact been standardized in ITU-T P.10

**Tools and Subjective Labs Setup** Have website collection of tools. Migrated to a new platform. If have tools for the community, reach out to this group.

**VQEG Administration and Web Suppor**t Reflectors continue to be a problem. No solution currently works for everyone. SAM has moved to Google Groups.

**MPEG** Working on point-cloud compression and neural network compression. Call for proposals on video compression beyond HEVC, SDR / HDR / 360 degree video.

**ITU-T SG12** Mentioned above (AVHD-AS/P.NATS Phase II) Also working on speech quality measurements, telepresence, P.QUITS initial buffering / rebuffering before people quit; how to perform subjective tests to detect the likelihood of people quitting, VR and 360 degree video (breakout today, and to be discussed later this week), , .

**Alliance for Open Media** AV1 has been  (late June / July); work on next version progressing; intense work on productizing, Microsoft indicates Windows / Chrome support, hardware being developed

**ICDM** Chair will come Wednesday to give presentation; working on next version of display standard, want to update chapter 4, visual inspection. ICDM is reaching out to VQEG for help with this update.

Presentations

Werner Robitza, “New Tools and Subjective Labs Website”

* New website based on GitHub to list available software, tools, and database.
* <https://vqeg.github.io/software-tools/>
* Look at the site, let Glenn and Werner know about new tools
* There is interest in merging the Qualinet database (http://dbq.multimediatech.cz/) into this website or a similar setup

**Action:** the group chairs will propose a collaborative VQEG / Qualinet effort to merge these two websites. A liaison will be sent from VQEG to Qualinet about it. Werner and Lukas will make a first draft

Computer Graphics (CGI)

Steven Schmidt, “Ongoing standardization activities of gaming quality of experience”

* ITU-T Rec. G.1032
* ITU-T Rec. P.809, includes subjective test methods (interactive vs passive)
* We are always searching for collaborations
* [www.qu.tu-berlin.de](http://www.qu.tu-berlin.de) for more information

Saeed Shafiee Sabet, “Gaming Video Datasets”

* Gaming Video dataset (KU-DT/TUB), KUGVD (KU), and CGVDS (DT/TUB-Simula)
* Dataset distributed soon
* Investigated usage rights from game owners: this use and redistribution is allowed

Saman Zadtootaghaj, “Performance evaluation of existing quality models and ITU standards on video gaming quality estimation”

* Cloud gaming: diverse in video complexity
* Existing metrics rarely trained no this content

No Reference (NR) Metrics  (NORM)

Mikołaj Leszczuk, “Quality Indicator - Lip Sync”

* Summary of lip sync influence, taken from various reports
* Metric to detect lip sync problems, through comparison of audio and video

Yilin Wang, “Quality analysis for user generated contents”

* Many videos uploaded to YouTube are not pristine, so full reference metrics do not work
* Described several NR metrics from published papers; references given
* Idea: report scores for each artifact + overall quality assessment

Nabajeet Barman, “ML based no reference gaming video quality estimation”

* Passive gaming video streaming
* Three models: NR-GVSQI, NR-GVSQE, NR-GVQM
* Datasets GamingVideoSet (available open source) and KUGVD
* Three different approaches to machine learning for NR metric design

Margaret Pinson, “What should an NR metric scatter plot look like?”

* Mylene Farias (Unb): Scores of different FR and NR metrics vary (in a good scenario, are scaled) for different datasets with different spatial and temporal resolutions.  In one previous work by our group, we tested the performance of metrics to these variations and adapted the videos (automatically) to obtain the best performance. Our goal in this case was actually to reduce both these dimensions to reduce computational complexity, but the same idea can be used in the design of NR metrics, since the same parameters are also available. One interesting point from this work is that the effect of spatial and temporal resolutions depends on the type of distortion, which means that a classification of the type of distortion is certainly useful.

(For details, see:  <https://www.researchgate.net/publication/327828616_A_framework_for_computationally_efficient_video_quality_assessment>)

* Lukas Krasula (Arm): The change in performance of particular sub-metrics could potentially be resolved by their fusion using a locally adaptive system, as proposed by Barri et al. (<https://ieeexplore.ieee.org/document/6786302>).
* Lukas Krasula (Arm): The task of evaluating quality of images/videos that are presented to observers only once is very challenging and will most probably result in high variance of observer scores. It is therefore dangerous to take MOS without considering their respective standard deviations. I would recommend the use of the performance evaluation methodology presented in here: <https://ieeexplore.ieee.org/document/7498936>. It not only considers standard deviations of the subjective scores but also enables meaningful combination of data coming from multiple databases without introducing noise caused by different quality scales among the experiments conducted when creating them. I am willing to provide any necessary assistance when processing the data using this method.

Discussion on No Reference Metrics

Where do we want to go? What do we want to accomplish?

Facebook; way of measuring user created content; willing to help with content generation

Limited to human consumption of video (computer use is separate project)

Professional & consumer generated content both of interest

Standardization only useful across industry (do not need internally); measurement standard

Sky UK: professional content, confidence score for the measurement (e.g., 4 mos +/- 1)

AT&T/DirecTV: professional content, contribution quality, distribution quality,

We can’t make it better but can make it much worse;

If incoming quality is below a threshold, then reject (if have standard scale, company can set a threshold)

AGH: What is quality? We don’t have a strict definition. This would help us limit scope (this part does not interest me vs this part does)

Map back to VMAF? Or include satisfaction, aesthetics, etc.?

Need a definition or example of the best possible quality.

QoE is slightly different from video quality. Enjoy watching pets even if poor quality. QoE includes audio quality, and other factors. If restrict to just video, still not one size fits all (e.g., live video vs gaming).

White paper / open metric vs standard?

* NR is too open / vague right now for an ITU Recommendation. ITU-T SG12 is discussing how to treat machine learning models.
* Some skepticism expressed on standardizing a list of numbers (machine learning)
* Qualinet white paper on QoE, see ITU-T Rec P.10. However, also some abuse exists.
* Separate standard from development work

Bar is low. Typical consumer equates resolution to quality.

To have impact, we must target a business problem, business decision

Artistic quality probably must be left out.

Standardize our viewing conditions. Device has impact, both in development & validation. Make these high level, practical decisions.

Original goal: toolsets to enable industry decisions. Consumers have difficulty understanding MOS from metrics (PSNR, VMAF). Need application specific tools that engineers understand and can use, including “**what is a meaningful difference**” for their specific application. Engineers don’t understand the numbers from our metrics.

JND is a solution to above, mentioned at a prior meeting.

**Decided:**

* Will start monthly calls.
* Goal of NORM:  Define a No-Reference metric that can accurately measure the visual quality of the decoded video “signal.”  Has at least a Mean Opinion Score (MOS) as a metric.  Has a specified just-noticeable-difference (JND) and confidence interval.  Has guidance on how to interpret the results.

**Open Actions:**

* Discuss MOS and JND on first call.
* Do we combine User Generated and Camera Capture?

**Professional Content:** Sky UK, Spirent, Facebook, AT&T/DirecTV, YouTube, Kingston Univ.

**User Generated Content:** YouTube, Facebook, Spirent, ByteDance, NTIA/ITS, Kingston Univ.

**Camera Capture:** NTIA/ITS (This may be merged with user generated content)

Day 2 – Tuesday, Nov 13

SAM (Statistical Analysis Methods)

Intro (Lucjan). SAM is working on:

* Subject model (tbd on Thursday)
* Other stuff: statistical analysis in general, format to report subjective tests, tc.

#104, Kjell Brunnström, “Statistical quality of experience analysis”

* Link to original article: <https://www.spiedigitallibrary.org/journals/journal-of-electronic-imaging/volume-27/issue-05/053013/Statistical-quality-of-experience-analysis--on-planning-the-sample/10.1117/1.JEI.27.5.053013.full>
* Within-subject design (typical scenario in subjective tests is that each user scores several PVSs): dependent T-test for paired samples.
* Estimation of sample size for statistical power (e.g. 0.8), MOS difference (0.5, 1.0), standard deviation (0.8), number of comparisons.



* Prepared input to ITU SG12: amend P.1401 with a text on multiple comparison, harmonization of texts in BT.500, P.910, P.913.
	+ No pre-planned number of comparisons: 27 in controlled env, 37 in public env.
	+ Pre-planned number on comparisons (<100): 18 and 25 subjects respectively.
* Comments about the proposal:
	+ Patrick: “magic” numbers in standards may be misleading; it’s better to propose tools
	+ Ioannis: other statistical analysis tools (e.g. removing user bias) may reduce the standard deviation and thus the number of subjects needed
	+ Further discussion on the topic (Margaret, Lucjan, Ioannis, Patrick).

#103, Lucjan Janoswki, [“JSON for Subjective Experiments”](https://docs.google.com/presentation/d/1QwbjwIv0mgbqor5EtARce1OvK0I_2UdOkS3dEVrGr1E/edit#slide=id.g47d2227892_3_49)

* Based on *sureal* package
* Per PVS, per subject, object-oriented
* Discussion on naming, etc. (several attendees)
* suJSON: a JSON for subjective tests’ data
* **Proposals:**
	+ **Add a field for pre- and post-experiment Q&A**
	+ **Add viewing conditions**
	+ **Prepare format for pair comparison (e.g. PVS\_id list at scores)**
	+ **Do not duplicate information (remove PVSs scored by subject in subject data, and subjects scoring PVSs in PVS data)**

IMG (Immersive Media Group)

#109, Francesca de Simone, “Measuring User Quality of Experience in Social VR Systems”

* User QoE in sVR has multiple dimensions (e.g. social presence, AV quality)
* Factors impacting QoE:
	+ Avatar appearance
	+ Avatar behavioural realism
* Experiment: view a movie trailer in pairs
	+ Compare: f2f, Facebook, TNO
	+ Measure:
		- Subjective: Quality of Interaction, Social Connectedness, Presence/Immersion
		- Objective: user behaviour (user talk to / look at each other)

#110, Ashutosh Singla, “Comparisons of Different Subjective Test Methods for HEVC Encoded Omnidirectional Videos”

* Measure user behaviour, video quality, simulator sequence
* Compare ACR, m-ACR, DSIS
* Results
	+ 4k vs 6k is observed, 6k vs 8k does not show any difference
	+ DSIS shows (slightly) better discriminative power than ACR
	+ Simulator sickness: increases along time (and with test duration), decreases with breaks, DSIS shows less sickness
* Discussion:
	+ Discussion on resolution of the display vs resolution of the video
* “TUIL 360º Video Dataset”
	+ 3x 8K 30-second high-quality video
* Preview of HVEI’19 presentation

#112, Pablo Perez, “Using 360 VR Video to Improve the Learning Experience in Veterinary Medicine University Degree”

* Preview of HVEI’19 presentation

QACoViA

#104, Mikołaj Leszczuk, “Evaluation of Multimedia Content Summarization Algorithms”

#105, Mikołaj Leszczuk, “Objective Video Quality Assessment Method for Recognition Tasks”

* Existing QoE objective metrics focus on “entertainment media”. They cannot predict the accuracy of recognition tasks with enough reliability.
* New project. Target: Develop a new objective video quality assessment method for recognition tasks.
* Evaluate against computer vision success instead of MOSs
* Comment (Patrick): What it is needed in QoCoVIA is establishing the statistical framework on how to test the quality estimator against the groundtruth. We also need a framework to “stress” the algorithm (which kinds of distortions / noise / … are applicable to CV use case).
* Spirent (Jeremy) is also interested in the topic. Today most data is transported uncompressed because of the fear that there might be issues if it is compressed. IEEE P2020 might be a group to collaborate with.
* Comment (Ioannis): is it possible to assign a score to a single signal (which will individually success or fail)?
* Mylene: should we enter into detecting images that can fool the algorithm? Patrick: we should not.

Immersive Media Group

#115, Kjell Brunnström, “Quality of Experience of Visual-Haptic interaction in a Virtual Reality simulator”

* Control of a truck crane with VR
* Test joystick delay in VR simulator
	+ 0 - 800 ms
* 31 test subjects (22m/9f), 2 didn’t finish the test due to sickness
* Results
	+ 0-400 ms no effect. 800 ms has significant effect in task completion (also subjectively).
* Next steps:
	+ Evaluate with experts (crane operators)
	+ Test with a remote controlled miniature crane, and also with the actual crane.

#129, Neil Birkbeck, Roberto Azevedo, “Understanding distortions in Youtube omnidirectional videos”

* Artifact ranking:
	+ Blur, blockiness, seams, h264 qp
	+ Focus: 1st JND and immersion breaking point
* Testing procedure:
	+ Find JND (user add degradation until notice difference with original)
	+ User add degradation until “immersion” is broken.
	+ No time limit per user
* Results
	+ Artifacts tend to be more annoying (break immersion) in stereo than in mono
	+ Existing metrics do not quantify localized artifacts (e.g. seams)
	+ Blurring seems to be less perceptible than blocking

JEG Hybrid

#108, Enrico Masala, “Objective Measures on the ITS4S Database”

* Objectively-annotated Large Scale Database
* Subjective annotation of ITS4S database
	+ ITS4S already contains subjectively annotated data
	+ Comparison with objective metrics (mostly VMAF)
* Data available for download

Florence Agboma, “Characterising Objective Metrics Using Large-Scale Database” (discussion session)

* Scope: Compare open source with commercial tools (objective metrics)
* Look for places where metrics agree and put the ground truth there (and do the subjective tests there).
* Understand weaknesses → better interpret the results

SAM / IMG

Lucjan Janowski, “Repetition button”

* Button where users report that they were unable to rate the sequence (they were distracted)
* Two options: repeat button vs skip (discard that subject’s score for one PVS)
* SAMVIQ allows re-play of videos
* Medical doctors allowed to replay because they are used to repeatedly viewing sequences
* Comparison tests with small differences are improved by repeat option
* Label? Maybe “I missed the sequence”
* Should note who used the button, for which sequences, who used the button often, …
* Self reporting & discard (skip) will have less impact on ACR task
* Could repeat at a random time later in the session (i.e., discard this score, come back to later)
* VQEG should consider using experts in subjective tests who can more consistently and accurately score video than consumers.

Day 3 – Wednesday, Nov 14

New Group Proposal

Proposal for new VQEG group: 5GKPI presented by Pablo    Perez

* Relationship between 5G key performance indicators and video QoE
* Proposal for starting group, in the future the group will propose test plans

Discussion:

* Is the group limited to 5G or any new network technologies?
	+ it is a good suggestion
* What part of 5G we are talking about? Is it the first step to answer this question by finding where are the problems?
	+ We start from entertainment video and self-driving cars.
* SG12 already decided to work on 5G. The work should be synchronized.
* Who is interested?
	+ NTIA/ITS,
	+ University of West Scotland,
	+ University of Madrid,
	+ Intel,
	+ Facebook,
	+ Opticom,
	+ Spirent,
	+ University of Nantes,
	+ Vittorio,
	+ Kingston University,
	+ RISE (Acreo)
	+ Nokia
* SG12 discussions, exchange of ideas would be helpful

IMG

Pablo provided summary of what was done.

Presentation: Narciso García    “Video Multimethod Assessment Fusion (VMAF) on 360VR”

*Discussion and Questions:*

- What is the input format for the VMAF

   - it is the rectangular form

- what was the subject behavior

  - they mostly watched equator

- It is surprising that VMAF works so well, for pools some distortions, not visible for VMAF should distort.

- Maybe the detail of the algorithm is limiting the pools influence

- only the same device?

  - yes

- how voting was made?

  - within an application by pressing a button on a headset

- Maybe you bias the chosen video by getting VMAF results before the experiment. There was a long discussion if the experiment design influenced the results. There are different ideas who it influenced the results.

- Did you use a different metric? If PSNR works as well then we know that it is not proved that VMAF is working so well.

  - It is planned but not yet done

- The results show that influence on quality almost does not depend on the content, which is surprising.

ITU-T Q13 Reporting

Kazuhisa Yamagishi

* What is the difference between G.QoE-VR  and P.360-VR
	+ G.QoE-VR describes only factors influencing and P.360-VR describes the test methodology.

Showing video impairments which should be detected by NR metrics produced by Sky

By Florence Agboma

IMG

Francesca De Simone “Evaluation of immersion and presence in VR”

*Discussion and Questions:*

* Francesca proposed to start collaborative work on the measuring and defining presence.
* How can we measure the time duration presence
	+ No ready solution
* Is there any other people interested:
	+ There are some people interested.
* What is the main reason to understand presence, will you prepare sequence which fit more to specific presence level?
	+ This is part of QoE, that is why we have to measure

Pablo Pérez, Jesús Gutiérrez “Methodologies for subjective quality evaluation of short and long 360-degree videos”

    *Discussion and Question:*

* Do you know if 20 is not just exploring the surrounding but not "seeing" the details?
	+ we are still exploring, so any literature can be helpful.
* We do not show a long sequence because then the content become important.
	+ It is true that time duration and interest in the sequence start to be important.
* People attention after 12 minutes drops significantly
	+ It is true, it is one of the problems related to this methodology.
* What are we asking about? Do we have MOS?
	+ no MOS in most of the cases, we ask more detail questions.
* Can we divide long sequence too short and model the head movements as the separate influencing factor?
	+ yes, but it will not cover all the use-cases.
* Reaction time will have a strong influence on continues voting.
	+ yes, it is a problem
* Do you assume that we are using TCP type service? ITU did the evaluation with one minute long sequences.
	+ This is a separate problem - quality fluctuations.
* Long sequences voted also divided into shorter sequences. It helped with the better understanding of the user behavior.
	+ it is a good idea.

Jesús Gutiérrez “Monitoring user behavior in subjective tests with 360-degree video”

*Discussion and Question:*

* For any experiment head movement, data should be gathered.
* We can use the evaluated metrics to predict which content should be used for subjective experiments.

ICDM Presentation

Joe Miseli, “ICDM standard on visual assessment, an often hidden gem”

* Version 2 coming, [www.icdm-sid.org](http://www.icdm-sid.org)
* ICDM is making HDR test videos and images that will be make these freely available
* Opportunity to contribute subjective methods to assess display performance
* 6 month window

PsyPhyQA

Psycho-physiology quality assessment, Neem, Sebastian, and Ulrich; Maria very active

* Summary of main  objectives

#128, Naeem Ramzan, Stamos Katsigiannis, “A QoE and Simulator Sickness Evaluation of a Smart-Exercise-Bike Virtual Reality System via User Feedback and Physiological Signals”

* Assess exercise game’s VR display, when varying resolution and frame rate
* Visual quality and simulator sickness

PsyPhyQA test plan is available at <https://docs.google.com/document/d/1o52KrvHhUCdrs4lGjWKtDjkPWrcaOof0kOLLXdXfR7E/edit>

* Have subjective plan ready; seeking other labs interesting in running this test
* UWS, Fraunhofer, Kingston University, and maybe Univ of Nantes

The next **eLetter** will be on PsyPhyQA topic. Announcement sent to reflector.

QoMEX 2018 had a session on PsyPhyQA.

Book to be published soon: **AI for Emerging Verticals**, subtitle: Robotics, human computing interaction, sensing and networking

Future Meetings

* Will seek a host in China and/or co-loc SG12 interim meeting
* 1 year out, possibly Sky UK, UPM, UWS, Kingston University
* Amazon would like to host, maybe fall 2019, Seattle USA

Presentations

#117, Yasuko Sugito, “8K subjective evaluation experiments”

* Describes challenges from 8K resolution monitors

#128, Pablo Perez, “Key Performance Indicators of Video and 5G networks”

#136, Mylene Farias, “Analyzing the influence of cross-modal degradations on the perceived audio-visual quality”

Day 4 – Thursday, Nov 15

SAM

134, Lucjan Janowski, “Notation proposal”

    *Discussion and Question:*

* It is suggested to use sigma i, sigma j,... instead of upsilon, phi, …
There is strong rejection considering this variable in the future could not be standard deviation
* Suggestion to reconsider i, j, k as definitions and use other letters
Off line meeting will be held to decide

101, Lucjan Janowski, “Maximum Entropy Solution”

    *Discussion and Question:*

* Clarification is asked about the choice of maximum entropy

119, Krzysztof Rusek, presented by Lucjan Janowski , “Towards Bayesian Subject Model”

113, Jing Li, “Hybrid-MST: A Hybrid Active Sampling Strategy for Pairwise Comparison”

    *Discussion and Question:*

* It is commented that parallel execution could reduce complexity

111, Zhi Li, Analysis Tools in the VMAF Open-source Package

* VMAF includes tools to train and understand model performance, particularly better understanding of machine learning models trained to predict quality
* New techblog, VMAF: the journey continues: <https://medium.com/netflix-techblog/vmaf-the-journey-continues-44b51ee9ed12>
* Packet Video Workshop 2019, June in MA, USA

*Discussion and Question:*

* Additional features for analysis are proposed which should already be in VMAF but not exposed yet.
* It is commented that the bootstrapped results do not mean around the result of the full model. More research is needed to explain this behaviour.
* It is suggested to not restrict bootstrapping to video or subject, but to sampling any user vote of a PVS. This strategy was tested, but feedback from statistical experts indicated that the variation would then be underestimated.

Presentations

139, Ioannis Katsavounidis, “The many faces of Video at Facebook”

*Discussion and Question:*

* Are there HDR plans?
Yes, as soon as the content is delivered to Facebook

138, Derek Pang, “Video quality considerations for Android”

*Discussion and Question:*

* It is commented that standards like P.1203 already take into account the different viewing modes mobile and TV. Also with respect to test data, there is already a lot of different types of content used these days. Content type contributions are encouraged from anyone having specific content requirements.
* Did you ever do test equipment using a high end camera as reference?
Controlled labs are used to perform tests.
For control, full-reference metrics are indeed already used.
* Will you be sharing any of your no reference metrics?
At the moment no end-to-end reference metric and no plan to generate one.
* Can subjective test result be made available?
Will be considered.
* Analyzing the users cannot be done because of privacy, could you setup such a system using volunteers at massive scale?
This has been actively looked at.
* Have you seen evidence that localized artefacts influence opinion scores?
Depending on the situation, this indeed can be bad.
* Couldn’t you analyze the camera performance on a camera basis rather than a photo basis?
The video quality rather than the camera quality is investigated.
Although sensor noise and capture noise is the same, the processing happening inside can be very different.
* Aren’t these operations under control of Android?
Indeed, but the API developer can do a lot of tuning.

#142, Ioannis Katsavounidis, “Video codec comparison using the dynamic optimizer framework”

-     Opportunities using FR metrics to choose best resolution, QP, and codec (e.g., H.264 vs H.265 vs VP9)

#140, Narciso García, “Considerations on FTV quality assessment”

* FTV achieved in real time

#130, Ashutosh Singla, “Approaches for assessing the simulator sickness”

*Discussion and Question:*

* Difference from simulator and cyber sickness. Relation with immersive media?

3D depth needs to be calibrated

* Questionnaire not designed for this use case. Users might be biased. Probably simulator sickness questionnaire should not to be included in recommendations.

Physiological information should be included (information is noisy)

HFVE (Human Factors for Visual Experiences) Session

#137, Sanghoon Lee, “IEEE 3D human Factor WG: current status of P3333.1” (Presented by Maria Martini)

* Related to immersive content (not only 3D)

#135, Maria Martini, “Quality of Experience for Light Field imaging”

* Report on three studies: Impact on angular resolution, image resolution, and joint impact

Contribution by IMG?

PsyPhyQA and SAM could also be involved

**Decision**: Collect proposals and contributions by VQEG and open a line of communication with the IEEE standardization group (e.g. by somebody reporting in that group)

Day 5 – Friday, Nov 16

IMG

Work on test-plan, Pablo Perez

* Use cases: discussion, if entertainment video is going to be the mass application of 360 video?
* Watching 360 video sequences:
	+ Some shooting conditions are not well-suited for 360 video.
	+ Still difficult (unsuccessful) shooting footages should be kept in a dataset.
* More information in the following document: <https://docs.google.com/document/d/18L_Q7vyzEssoAzbRsTKBTtMrwfjISMf0HUGofPqOOJI>

#110, Ashutosh Singla, “Comparisons of Different Subjective Test Methods for HEVC Encoded Omnidirectional Videos” (fragments repeated, discussion on stitching artifacts)

Jesús    Gutiérrez, 360 content datasets

* More information in the following document: https://docs.google.com/document/d/1Q23TI2ARHWjP\_13Ngya9pp672CCo1WaK1wHI4eCCRc8/edit
* Discussion on subjective test methodology

Agreements:

* Test-plan for subjective test methodology to be developed
* Cross-lab common set to be created
* Test-plan by the next meeting
* Open questionnaire at the end of the test to capture open thoughts of subjects
* **Decision**: We will be testing:
	+ Length of sequences
	+ ACR vs DCR
	+ Influence of HMD
	+ Simulator sickness questionnaire or other relevant questionnaires
	+ Random viewports (To be decided)
* Collecting partners interested in conducting experiments:
	+ YouTube
	+ Gent
	+ RISE
	+ Evaltech
	+ UPM
	+ Ilmenau
	+ Nantes
	+ Nokia Bell Labs
	+ Kingston
* Interested partners from IMG meetings (not present)
	+ CWI
	+ Roma 3